Du modèle génétique au modèle statistique

Vincent Ducrocq

V. DUCROCQ

INRA Station de Génétique quantitative et appliquée 78352 Jouy-en-Josas Cedex

Les bases de la génétique quantitative

Du modèle génétique au modèle statistique

Résumé. *L'étape préliminaire à l'étude d'un caractère quantitatif est la modélisation statistique des observations, c'est-à-dire la représentation mathématique d'une réalité biologique, cohérente avec le modèle génétique. Ce modèle décrit très généralement chaque observation comme étant la somme d'effets du milieu et d'effets génétiques. Différentes façons de modéliser chacune de ces parties sont possibles en fonction des objectifs de l'étude, et ont des conséquences diverses sur les hypothèses requises pour l'obtention de résultats fiables. Ainsi, il est essentiel d'inclure dans l'analyse uniquement les facteurs du milieu ayant un effet important sur la performance observée, mais sans en omettre aucun. On peut ne faire intervenir dans la partie génétique du modèle que des animaux apparentés aux animaux "auteurs" des observations. Des simplifications importantes notamment calculatoires peuvent en découler, mais les résultats peuvent alors être fortement biaisés si certaines hypothèses qu'il faut faire ne sont pas vérifiées. Enfin, selon qu'une connaissance a priori des caractéristiques statistiques des effets est disponible et utilisée, il est possible de distinguer trois types de modèles (à effets fixés, aléatoires ou mixtes) auxquels correspondent des méthodes différentes d'estimation des paramètres étudiés.*

L'étude d'un caractère quantitatif à déterminisme génétique supposé polygénique utilise systématiquement deux sources d'information : d'une part, des observations ou "performances", mesurées dans une population donnée, d'autre part des généalogies (connaissance des filiations). Pour étudier les mécanismes de sa transmission héréditaire, on est contraint de passer par une représentation mathématique de la réalité : c'est la modélisation. Le modèle statistique que l'on utilisera pour décrire les observations sera du même type dans le cadre d'une estimation des paramètres génétiques (héritabilités, corrélations génétiques avec d'autres caractères) que pour une évaluation de la valeur génétique des reproducteurs ou l'analyse de plans d'expérience et devra posséder deux caractéristiques fondamentales :

- il devra être opérationnel, c'est-à-dire qu'il devra être suffisamment proche de ce que l'on connaît de la réalité biologique ou des conditions d'élevage pour ne pas fausser l'analyse, tout en restant suffisamment simple pour obtenir des résultats fiables et sensés. Comme on le verra, le perfectionnisme dans la définition des modèles peut conduire soit à des impossibilités d'estimation, soit à des résultats erronés.

- il devra être cohérent avec le modèle génétique, c'est-à-dire qu'il devra explicitement prendre en compte nos connaissances (ou nos hypothèses) sur le déterminisme génétique du caractère et en particulier les ressemblances entre performances d'individus apparentés.

1 / La modélisation

Le modèle de base qui tente d'expliquer la "genèse" d'une performance est toujours de la forme :

P = G + E (1)

performance = effet du génotype + effet du milieu

Génotype et milieu sont supposés agir de façon indépendante (lorsque ce n'est pas le cas, on parle d'interaction génotype x milieu (G x E) et les choses se compliquent. En général, on considérera alors que des performances réalisées dans des milieux différents sont des caractères différents).

1.1 / P = la performance

En fait, P est rarement une observation élémentaire mais déjà une variable "synthétique", standardisée, précorrigée. En effet, ce qui est observé sur le terrain, ce sont des kilos de lait à la traite, des poids à un âge donné, la présence ou l'absence d'un oeuf un jour donné. Or, en pratique, toujours dans le but d'être "opérationnel", on se contente de résumer ces données élémentaires en une production totale par lactation, un gain moyen quotidien ou un nombre total d'oeufs pondus. Puis, si besoin est, on les standardise, afin de corriger d'éventuels effets d'échelle. Par exemple, on calcule une production laitière "standard" en 305 jours et ramenée à un niveau adulte (correction multiplicative pour le numéro de lactation, Bonaïti *et al* 1990).

Notons d'autre part qu'à un même caractère étudié peuvent correspondre plusieurs mesures différentes :

à titre d'exemple, la fertilité d'une femelle peut être mesurée par son intervalle entre mise bas, son intervalle entre la 1ère saillie et la saillie fécondante, un code de réussite à l'insémination (0 = échec, 1 = succès) défini soit à la mise bas, soit par un non retour en chaleur n jours après celle-ci, etc....

1.2 / E = les effets du milieu

Une modélisation correcte des effets du milieu est tout à fait déterminante : il ne serait pas raisonnable d'envisager une étude et des méthodes très sophistiquées sur des données décrites selon un modèle d'analyse déficient.

Dans cette partie "effets du milieu", on peut distinguer :

- Un effet de milieu particulier : la moyenne générale de la population.

- Les effets du milieu identifiés et enregistrés, dont on pense qu'ils ont une influence sur la performance. On sait par exemple que le mois et l'année de vêlage d'une vache, son âge au vêlage, son élevage (qui caractérise une alimentation et un type de conduite du troupeau particuliers) conditionnent son niveau de production laitière.

- Les effets identifiés mais non enregistrés ou non enregistrables tels que - toujours dans le cas de la production laitière - l'état reproductif de l'animal (gestant ou non), le temps séparant 2 traites, etc... qui ont un impact certain sur la performance mais ne peuvent être pris en compte explicitement car il s'agit d'une information trop coûteuse ou trop difficile à recueillir.

- Les effets non identifiés - tels que les conditions particulières lors de la mesure (état sanitaire de l'animal, place dans la salle de traite) - ou non identifiables (erreur sur la mesure élémentaire ou sur le calcul de la performance standard à partir des mesures élémentaires). On supposera que ces deux dernières catégories incluent de nombreux facteurs ayant chacun un effet faible sur la performance. Nous verrons plus loin les conséquences d'une violation de cette hypothèse.

En fait la modélisation va consister à retenir parmi les effets du milieu identifiés et enregistrés tous ceux (mais uniquement ceux-là) qui jouent un rôle essentiel - "significatif", avec tout ce que ce terme contient d'arbitraire - dans l'élaboration de la performance.

La description de ces effets dans le modèle se fera à travers le choix d'une relation mathématique décrivant la nature de la liaison entre la performance et une variable - appelée covariable - caractérisant l'effet en question. Par exemple, on écrira que l'effet de l'âge sur le poids de l'animal est décrit par la relation :

$$y = \mu + \alpha_1 x_1 + \alpha_2 x_1^2 \text{ (+ autres effets)} \quad (2)$$

où y est le poids d'un animal de l'élevage i et x_1 est la covariable "âge de l'animal en mois" et α_1 et α_2 sont des "coefficients de régression".

Lorsque l'on ne souhaite pas ou que l'on ne peut pas préciser la nature de cette liaison - quelle covariable utiliser pour décrire au mieux l'effet élevage ? - on définira des classes (ou niveaux) d'un facteur décrivant cet effet (le facteur "élevage") et on écrira par exemple :

$$y = \mu + h_i \text{ (+ autres effets)} \quad (3)$$

avec h_i : effet du ième niveau du facteur "élevage"

Ces deux types d'approche sont tout à fait équivalents. En effet, si pour chaque observation et chaque niveau du facteur élevage, on définit une (co)variable x_i dite "indicatrice" telle que $x_i=1$ si l'observation "y" a été obtenue dans l'élevage i et $x_i = 0$ dans le cas contraire, on peut écrire alors d'une façon générale :

$$y = \mu + h_1x_1 + h_2x_2 + ... + h_ix_i + ...+ h_n x_n \text{ (+ autres effets)} \quad (4)$$

Ce type de modèle où une observation est décrite mathématiquement comme combinaison linéaire de paramètres à estimer (les α_i ou h_i) est appelé modèle linéaire.

1.3 / G = effet du génotype

Cette partie dépend essentiellement du type d'analyse que l'on veut effectuer. Pour l'illustrer, nous décrirons brièvement 4 exemples de situations courantes : évaluation de reproducteurs sans ou avec effets maternels, estimation de paramètres génétiques, analyse de plans de croisement.

a / Evaluation des reproducteurs pour un caractère non soumis à des effets maternels

(Henderson 1984, Quaas 1984, Ducrocq 1990).

On peut séparer l'effet du génotype en la somme d'un effet génétique additif (ou "valeur génétique additive") égale à la somme des effets moyens des gènes, et d'une valeur génétique non additive, somme des interactions - dominance et épistasie - entre gènes. En pratique, on ne s'intéresse le plus souvent qu'à la valeur génétique additive "a_j" d'un reproducteur j car c'est celle qu'il est susceptible de transmettre à sa descendance. Par opposition, la partie non additive du patrimoine génétique est recréée aléatoirement à chaque génération et n'est donc pas directement utilisable dans les schémas de sélection.

L'équation du modèle s'écrira, pour une observation l utilisée afin d'évaluer un animal j :

$$y_l = g_j + (\mu + \sum f_{(l)}) + e_l \quad (5)$$

où :

g_j représente l'effet génétique de j sur la performance l.

$\sum f_{(l)}$ est la somme des effets du milieu identifiés influençant la performance y_l.

e_l est la résiduelle ou l'erreur du modèle et englobe tout ce qui n'a pas encore été pris en compte, c'est-à-dire les effets de milieu non enregistrés et non identifiés, mais aussi la partie génétique non additive.

Il est important de noter ici que l'observation l peut avoir été obtenue sur un animal différent de (mais apparenté à) l'animal j, par exemple sa fille et alors $g_j = (1/2)a_j$. Lorsque j est l'animal auteur de la performance mesurée, le modèle est appelé modèle individuel ou modèle animal et on a $g_j = a_j$.

b / Evaluation des reproducteurs pour un caractère soumis à des effets maternels

(Henderson 1984 Quaas 1984)

Certains caractères - tels que le poids d'un veau à la naissance ou au sevrage - sont contrôlés à la fois par les gènes du veau j (on parle d'effets génétiques (additifs) directs a_j) et les gènes de sa mère m (effets génétiques maternels) t_m. Ces deux types d'effets doivent être considérés simultanément dans la partie "influence du génotype" sur la performance P et l'équation d'un modèle animal avec effets maternels s'écrira pour une observation l=j, de l'animal j dont la mère est m :

$$y_l = (a_j + t_m) + (\mu + \Sigma f_{(l)}) + e_l \qquad (6)$$

(Il est à remarquer que les effets "du milieu" peuvent ici inclure des caractéristiques non génétiques de la mère, telles que son âge par exemple).

c / Estimation des paramètres génétiques dans un schéma d'accouplement hiérarchique

(Minvielle 1990)

Dans certaines espèces, en particulier pour la volaille ou le lapin, les accouplements sont très hiérarchisés. (Au moins en théorie) chaque mâle est accouplé à n femelles et pour chaque couple, on mesure k descendants. Classiquement, le modèle d'analyse pour l'évaluation des reproducteurs ou pour l'estimation des paramètres génétiques s'écrira alors :

$$y_{ijl} = (p_i + m_{ij}) + (\mu + \Sigma f_{(l)}) + e_{ijl} \qquad (7)$$

avec i=1, ... n_p (nombre de pères) j=1,... n et l=1, ... k.
p_i est la contribution du père i à la performance de son lème descendant issu de la jème mère à laquelle il a été accouplé.

m_{ij} est la contribution de la jème mère accouplée au mâle i à la performance de l.

Cette "hiérarchisation" a des conséquences importantes : la contribution du père est purement d'ordre génétique additive ($p_i = (1/2)\, a_i$) alors que la contribution de la mère incluera à la fois l'effet génétique additif, les effets génétiques non additifs (dominance, épistasie) propres au couple i,j et l'effet maternel de la mère j sur la performance de l.

d / Analyse de plans de croisement

(Sellier 1982)

Dans l'analyse de plans de croisement, on ne s'intéresse plus en tant que tel à chaque individu mesuré mais aux races dont ils sont issus. Dans un dispositif diallèle par exemple, on compare les performances de tous les croisements possibles entre r races prises 2 à 2. Dans l'analyse de telles données, l'équation du modèle s'écrira par exemple :

$$y_{ijl} = (AG_i + AG_j + AS_{ij}) + (\mu + \Sigma f_{(l)}) + e_{ijl} \qquad (8)$$

où AG_i est l'effet génétique (appelé "aptitude générale au croisement") de la race i sur la performance de l'animal l, produit du croisement des races i et j et AS_{ij} est l'effet (appelé "aptitude spécifique au croisement") de la combinaison des deux races i et j. Là encore, AG_i peut être interprété comme la contribution génétique additive de la race i et AS_{ij} comme la contribution génétique non additive du croisement i x j (= l'écart à la somme $AG_i + AG_j$).

1.4 / Reparamétrisation

Sans changer en aucune façon les caractéristiques essentielles de ces modèles, on peut modifier légèrement la description de certains effets. Par exemple, pour une observation de poids y, la partie "somme des effets de milieu" peut s'écrire :

moyenne générale + effet de l'élevage en écart à la moyenne générale + effet de l'âge.

ou bien :

effet de l'élevage (implicitement en écart à la valeur "0") + effet de l'âge.

Ces deux descriptions sont tout à fait équivalentes d'un point de vue statistique. Lors du passage de l'une à l'autre, on parle de "reparamétrisation". Elles sont malgré tout différentes d'un point de vue calculatoire : la première comporte en apparence une inconnue de plus : la moyenne générale (en apparence seulement, car en précisant que l'effet de l'élevage est exprimé en écart à la moyenne générale, on introduit une contrainte supplémentaire : on force les inconnues à respecter une relation particulière).

2 / Effets fixes, effets aléatoires

Sous l'hypothèse d'un déterminisme polygénique, les gènes influençant le caractère considéré sont supposés très nombreux et ayant chacun un effet faible. En application de la loi des grands nombres, la somme a_j des effets moyens de ces gènes suit une distribution normale de variance σ_a^2 (variance génétique additive). Nous connaissons également la covariance entre valeurs génétiques additives d'animaux apparentés, qui pour 2 demi-frères (ou demi-soeurs) s'écrira (1/4) σ_a^2 et pour 2 plein-frères (ou pleine-soeurs) $(1/2)\sigma_a^2+(1/4)\sigma_d^2$.... En fait, c'est l'ensemble des effets génétiques qui suivent une loi normale en plusieurs dimensions (une loi "multinormale") pour laquelle l'ensemble des variances et covariances sont des multiples de σ_a^2 que l'on regroupe dans une matrice, la "matrice de parenté".

Donc, avant même d'observer les performances, on connaît certaines caractéristiques de la distribution des effets à estimer et nous allons prendre en compte cette connaissance dans l'analyse des performances. On dit alors que la valeur génétique additive a_j est un effet aléatoire dans notre modèle. Cette terminologie est liée à la notion de variable aléatoire en statistique : une variable aléatoire est une variable qui peut prendre différentes valeurs avec une certaine probabilité associée à chaque valeur. La réalisation de la variable aléatoire "valeur génétique additive" de l'animal j peut être assimilée au résultat d'un tirage aléatoire d'un individu dans une distribution "normale".

De même, les facteurs d'origine génétique ou liés au milieu qui ne sont pas pris en compte explicitement dans notre modèle sont supposés nombreux et à effet faible : la résiduelle e_l est un effet aléatoire qui suit une distribution normale de moyenne nulle, de variance σ_e^2 (variance "résiduelle") et indépendante de la distribution de a_j.

Par contre, pour la plupart des effets du milieu, ou pour certains effets génétiques comme les aptitudes aux croisements AG_i, AS_{ij} , on ne fera aucune hypo-

thèse concernant leur distribution statistique : il est difficile d'imaginer que l'effet de l'âge, de la saison ou de la race sur le poids d'un animal est la réalisation d'une variable aléatoire obtenue par "tirage" dans une population d'effets "âge", d'effets "saison" ou d'effets "race" ayant une distribution normale (ou toute autre distribution). On parlera alors d'effets fixes ou fixés. La distinction entre effets fixés et aléatoires paraît souvent arbitraire : un effet "élevage" résumant l'ensemble des effets propres à cet élevage (microclimat, nature des sols, alimentation, conduite du troupeau, etc...) a plus les caractéristiques d'un effet aléatoire (=somme d'effets nombreux et de faible amplitude) que celles d'un effet fixe. Pourtant, dans la plupart des cas, pour certaines raisons non développées ici, on le considère comme fixe dans nos analyses. L'école statistique bayésienne (Berger 1985) propose une autre interprétation pour résoudre ce paradoxe : en caricaturant à l'extrême, tous les effets sont des variables aléatoires mais dans un cas - pour les effets fixes - on ne base l'analyse que sur les observations et dans l'autre - pour les effets aléatoires - on combine l'information apportée par les données à une information a priori sur la distribution des effets à estimer.

Cette distinction entre effets fixés et effets aléatoires conduit à la définition de 3 grandes familles de modèles linéaires :

2.1 / Les modèles à effets fixés

Dans ces modèles, tous les effets sauf la résiduelle sont des effets fixes. L'exemple le plus connu de tels modèles est la régression, simple ou multiple. La méthode classique pour l'estimation des paramètres d'un modèle à effets fixés est la méthode des "moindres carrés" pour laquelle on cherche à minimiser la somme des carrés des résiduelles. Par exemple, pour la régression simple de y sur x, si l'on a n observations y_i décrites par le modèle $y_i = \alpha_0 + \alpha_1 x_i + e_i$, les estimées des moindres carrés de α_0 et α_1 seront les valeurs qui minimisent :

$$\sum_{i=1}^{n} e_i^2 \quad \text{c'est à dire} \quad \sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 x_i)^2$$

2.2 / Les modèles aléatoires

Dans ces modèles, tous les effets sont aléatoires, sauf éventuellement la moyenne générale μ. C'est une situation classique en sélection avicole : l'ensemble des effets du milieu peut être résumé par un effet "lot" ou "groupe de contemporains élevés dans un même bâtiment". Comme à cet effet correspond un très grand nombre d'animaux (plusieurs centaines), on peut considérer qu'il est très bien connu et estimé par la moyenne μ_k des performances du lot k et on écrira :

$$(y_l - \hat{\mu}_k) = a_j + e_l \qquad (9)$$

si l'animal a réalisé sa performance l dans le lot k, où a_j a une distribution normale de moyenne 0 et de variance σ_a^2 et e_l a une distribution normale de moyenne 0 et de variance σ_e^2.

La méthode d'estimation de a_j fait alors appel à la théorie des index de sélection (Smith 1936, Hazel 1943) qui permet de combiner 2 sources d'information : dans le cas du modèle (9) et en supposant les animaux à évaluer non apparentés, on combine :

- une estimation obtenue à partir de la performance propre de l'animal

$$j : \hat{a}_j^{(1)} = (y_l - \hat{\mu}_k)$$

avec une "erreur" e_l et une "variance d'erreur" σ_e^2.

- une estimation obtenue à partir de la connaissance a priori de la distribution de a_j, qui rappelons-le est supposée normale, de moyenne nulle, et de variance σ_a^2 :

$\hat{a}_j^{(2)} = 0$ (la moyenne) avec une "variance d'erreur" σ_a^2

Et en calculant une moyenne de ces deux estimées, pondérées par l'inverse de la "variance d'erreur" connue sur chacune d'elle, on obtient la valeur génétique estimée $\hat{a}_j$ ou "index" de j :

$$\hat{a}_j = \frac{1}{\frac{1}{\sigma_e^2} + \frac{1}{\sigma_a^2}} \left(\frac{1}{\sigma_e^2} a_j^{(1)} + \frac{1}{\sigma_a^2} a_j^{(2)} \right)$$

$$\hat{a}_j = \frac{\sigma_a^2 \sigma_e^2}{\sigma_a^2 + \sigma_e^2} \left(\frac{1}{\sigma_e^2} (y_l - \hat{\mu}_k) + \frac{1}{\sigma_a^2} .0 \right)$$

$$\Rightarrow \quad \hat{a}_j = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} (y_l - \hat{\mu}_k) = h^2 (y_l - \hat{\mu}_k)$$

où on retrouve h^2, l'héritabilité du caractère.

Si les animaux mesurés sont apparentés ou encore si notre but est d'évaluer un animal apparenté à ceux qui sont auteurs des performances (leur père par exemple), les calculs sont plus complexes mais la démarche est la même : les performances sont d'abord corrigées pour les effets de milieu en supposant ceux-ci connus parfaitement. Puis on combine de manière adéquate informations a priori et informations apportées par les données.

2.3 / Les modèles mixtes

Les modèles mixtes qui combinent à la fois des effets fixes et des effets aléatoires. C'est le cas le plus général pour l'évaluation des reproducteurs. En effet, pour la plupart des espèces, on ne dispose pas pour chaque effet de l'information suffisante pour bien l'estimer seul et "a priori". Le fait de considérer les deux types d'effets (du milieu et génétiques ou fixes et aléatoires) simultanément permet avec une méthode de statistique d'estimation adéquate de corriger les estimées d'un effet pour tous les autres effets, qu'ils soient fixes ou aléatoires. Dans le cas des modèles mixtes, cette méthode statistique est le BLUP (Best Linear Unbiased Prediction ou Meilleure Prédiction Linéaire non Biaisée ; Henderson 1963, 1973) qui peut être regardée comme une utilisation de la théorie des index de sélection pour calculer des valeurs génétiques en fonction des données corrigées pour les effets fixes. Les estimées des effets fixes utilisées pour la correction sont - en théorie - obtenues par la méthode des moindres carrés généralisés. Cette der-

nière est une extension de la méthode des moindres carrés décrite plus haut, qui intègre le fait que les performances peuvent être réalisées par des animaux apparentés et peuvent donc être corrélées. Le BLUP possède des propriétés très intéressantes dont la plus importante est l'absence de biais dans l'estimation des effets aléatoires (à condition que le modèle de départ soit correct).

3 / Retour sur les facteurs du milieu

Que se passe-t-il lorsqu'on ignore dans notre modèle un facteur du milieu qui a un effet important sur les observations, par exemple le facteur "mois de vêlage" en production laitière ? La première conséquence de cette omission est l'estimation biaisée des autres effets et en particulier des valeurs génétiques : par exemple on imputera la plus faible production laitière d'une vache vêlant en été à une valeur génétique moins élevée plutôt qu'au climat. L'existence d'animaux pénalisés et d'autres avantagés aboutira à des erreurs de classement sur leur valeur génétique estimée (index) et donc à une baisse de l'efficacité globale de la sélection. Une autre conséquence de ces biais est de surestimer la précision de notre évaluation : celle-ci est en effet calculée sous l'hypothèse d'absence de biais systématique dans l'évaluation.

Inversement, l'inclusion dans l'analyse de facteurs du milieu inutiles, c'est-à-dire dont les effets sont sans importance réelle, est à éviter : elle ne biaise pas l'estimation des autres effets mais diminue la précision de leurs estimées : si, au lieu de l'effet "mois de vêlage" cité plus haut en production laitière, on introduit un effet "jour de vêlage", on ne décrira qu'un tout petit peu mieux l'influence du changement de conditions climatiques au cours du temps sur la production laitière mais pour une année donnée, on aura 365 paramètres à estimer au lieu de 12, à partir de la même quantité d'informations (performance et généalogie). On aura certes l'impression d'une meilleure adéquation du modèle (R^2 plus élevé) mais les estimées des effets "jour de vêlage" étant chacune très imprécises, la qualité prédictive du modèle sera plus faible. De plus, on accroît ainsi les risques de disconnexion dans l'estimation.

La disconnexion est l'absence d'information permettant une comparaison juste entre animaux. Elle entraîne une confusion entre effets. Par exemple (figure 1), supposons deux taureaux non apparentés A et B. Si A a n_A filles dans un troupeau 1 et B a n_B filles dans un troupeau 2, on ne sera pas en mesure de savoir si l'écart de performances entre les filles de A et celles de B est dû à une meilleure valeur génétique de l'un des taureaux ou à un meilleur effet "élevage". Il y a disconnexion lorsque le modèle d'analyse est de la forme $y_{ijl} = \mu + h_i + p_j + e_{ijl}$ où h_i est l'effet de l'élevage i (= 1 ou 2) et p_j est l'effet du taureau j (= A ou B). Par contre, si le taureau A a des filles à la fois dans le troupeau 1 et dans le troupeau 2, on pourra statistiquement estimer la part de chacun des effets sur les différences de performances observées : les filles du taureau A servent de connexion. (De même, si A et B sont demi-frères, la connaissance de leur apparentement apporte une information - mais moins "riche" que dans le cas précédent - sur la part des différences observées qui est d'origine génétique : les relations de parenté entre individus sont des sources de connexion).

Figure 1. *Illustration d'un dispositif connecté ou disconnecté.*

Disconnexion

		Nombre de filles dans le troupeau	
		1	2
Taureau	A	n_A	-
	B	-	n_B

Connexion

		Nombre de filles dans le troupeau	
		1	2
Taureau	A	n_{A1}	n_{A2}
	B	-	n_B

Le dernier élément à considérer dans le choix des effets de milieu est l'existence fréquente d'interactions. L'effet du mois de vêlage, par exemple, sur la production laitière n'est pas nécessairement le même pour une année humide que pour une année sèche. On définit alors souvent cet effet du mois de vêlage pour chaque année et on parle d'interaction "mois de vêlage x année". Mais là encore, la prise en compte d'interactions peut accroître les risques de disconnexion : si le taureau A a des filles dans les troupeaux 1 et 2 en 1989 et le taureau B, non apparenté à A, a des filles dans le troupeau 2 en 1990, un modèle d'analyse incluant un effet élevage x année sera disconnecté, alors que ça ne sera pas le cas pour un modèle incluant les effets élevage et année séparément (élevage + année).

4 / Retour sur la partie génétique du modèle (Quaas 1984)

Dans la variante "modèle animal" de l'équation du modèle (5), la valeur génétique additive considérée $g_j=a_j$ est celle de l'animal j, auteur des performances. La valeur a_j est la somme des effets moyens des gènes portés par les gamètes que j reçoit de son père p et de sa mère m respectivement. En espérance - c'est-à-dire "en moyenne sur un grand nombre de situations identiques" - la somme des effets moyens des gènes portés par le gamète provenant du père p de j est (1/2) a_p. Mais en espérance seulement, car les hasards de la ségrégation mendélienne et des recombinaisons au cours de la méiose ont pu donner à j plus de gènes favorables ou défavorables du père p que la moyenne des descendants de p. Le même phénomène a lieu

également pour le gamète provenant de la mère et on pourra écrire :

$$a_j = \frac{1}{2} a_p + \frac{1}{2} a_m + \phi_j \qquad (10)$$

où ϕ_j est la somme des écarts à l'espérance pour les gamètes du père et de la mère. ϕ_j est appelé l'"aléa de méiose" et est bien entendu indépendant de a_p et de a_m, les valeurs génétiques additives des parents. On peut montrer que ϕ_j est une variable aléatoire qui a pour variance :

$$\frac{1}{2}\left(1 - \frac{F_p + F_m}{2}\right)\sigma_a^2$$

où F_p et F_m sont les coefficients de consanguinité des parents.

De cette équation résultent quelques simplifications possibles dans notre modèle animal (5). Par exemple, si les parents p et m n'ont jamais de performances propres (ou bien si on choisit d'ignorer celles-ci), on pourra réécrire :

$$y_l = g_p + g_m + (\mu + \Sigma f_{(l)}) + e_l^* \qquad (11)$$

où : $g_p = u_p = \frac{1}{2} a_p$, $g_m = u_m = \frac{1}{2} a_m$, $e_l^* = e_l + \phi_j$

$$var(u_p) = \frac{1}{4}\sigma_a^2 \qquad var(u_m) = \frac{1}{4}\sigma_a^2$$

$$var(e_l^*) = \sigma_e^2 + \frac{1}{2}\sigma_a^2 = \left(\sigma_y^2 - \frac{1}{2}\sigma_a^2\right)$$

Un tel modèle s'appelle "modèle père-mère" et u_p, u_m sont les "aptitudes transmises" des animaux p et m. L'intérêt de cette simplification est qu'au lieu d'avoir un nombre de valeurs génétiques à estimer égal au nombre total d'animaux, on n'en a plus que n_p(nombre de pères) + n_m(nombre de mères). Les calculs s'en trouvent allégés si notre but est d'évaluer simplement les animaux parents.

On peut pousser la simplification plus loin en écrivant dans (10) a_m sous la forme :

$$a_m = \frac{1}{2} a_{gpm} + \frac{1}{2} a_{gmm} + \phi_m = u_{gpm} + u_{gmm} + \phi_m$$

(avec gpm = grand père maternel, gmm = grand mère maternelle et ϕ_m est l'aléa de méiose pour l'animal m) et en supposant les mères apparentées uniquement à travers leur père (les grands mères maternelles sont non apparentées et non sélectionnées), on aura le "modèle père-grand père maternel" suivant :

$$y_l = \left(u_p + \frac{1}{2} u_{gpm}\right) + (\mu + \Sigma f_{(l)}) + e_l^{**} \qquad (12)$$

$$\text{avec } e_l^{**} = e_l^* + \frac{1}{4} a_{gmm} + \frac{1}{2}\phi_m$$

ou bien, dans l'hypothèse extrême où les mères des individus mesurés ne sont pas apparentées, sont de valeur génétique moyenne nulle et n'ont jamais plus d'un descendant chacune, on peut définir un "modèle père" :

$$y_l = u_p + (\mu + \Sigma f_{(l)}) + e_l^X \quad \text{avec } e_l^X = e_l + \frac{1}{2} a_m + \phi_m \qquad (13)$$

Dans ces deux derniers modèles, seuls les parents mâles sont évalués : la taille du système à résoudre est considérablement réduite (et était la seule qu'il était possible de traiter avec nos ordinateurs pour les grandes espèces domestiques jusqu'à récemment). Mais il est absolument essentiel de remarquer que la contrepartie de cette réduction est la formulation d'hypothèses fortes, qui, lorsqu'elles ne sont pas vérifiées (et elles le sont rarement), peuvent entraîner des biais importants dans l'évaluation. En fait, dans cette cascade de modèles, la description de la partie génétique est de moins en moins précise : une fraction de plus en plus importante des effets génétiques se retrouve dans la partie non expliquée (la résiduelle) du modèle, comme le montre le tableau suivant :

Tableau 1 .*Caractéristiques de différents modèles génétiques.*

Modèle	Coefficients de a_i pour i = animal	Père	Mère	GPM	Variance de la partie génétique	Variance de la résiduelle
Animal (5)	1				σ_a^2	σ_e^2
Père-Mère (11)		0,5	0,5		$(1/4+1/4)\sigma_a^2$	$\sigma_e^2+1(1/2)\sigma_a^2$
Père-GPM (12)		0,5		0,25	$(1/4+1/16)\sigma_a^2$	$\sigma_e^2+(11/16)\sigma_a^2$
Père (13)		0,5			$(1/4)\sigma_a^2$	$\sigma_e^2+(3/4)\sigma_a^2$

Pendant longtemps, l'utilisation du modèle le plus précis - le modèle animal - n'était pas envisageable, pour des raisons de coûts informatiques principalement. Avec l'apparition de matériel sans cesse plus performant, il a été possible dans la plupart des espèces domestiques évaluées en routine de passer du modèle père au modèle père-grand-père puis enfin au modèle animal, qui utilise pratiquement au mieux l'ensemble de l'information disponible.

5 / Situations plus complexes

(Quaas 1984)

Il arrive que l'on souhaite étudier les aspects génétiques de plusieurs caractères corrélés simultanément. Dans ce cas on aura recours à une "analyse multicaractères". Le modèle statistique utilisé sera obtenu simplement en généralisant les modèles (5), (11), (12) ou (13). Par exemple, pour un modèle animal et 2 caractères y_{i1} et y_{i2} mesurés sur l'animal j, on écrira :

$$y_{l_1} = a_{j_1} + (\mu_1 + \Sigma f_{(l)_1}) + e_{l_1}$$

$$y_{l_2} = a_{j_2} + (\mu_2 + \Sigma f_{(l)_2}) + e_{l_2} \qquad (14)$$

où les indices 1 et 2 se rapportent à chacun des caractères et les notations sont inchangées par rapport à (5). Les effets fixes pour les caractères 1 et 2 peuvent être définis de manière identique ou de façon radicalement différente. La seule différence importante avec l'analyse unicaractère est que dans la distribu-

tion des effets aléatoires (valeurs génétiques additives et résiduelles), on fera intervenir toutes les covariances (génétiques et résiduelles) non nulles entre les performances sur chaque caractère pour un même individu ou pour deux individus apparentés. De cette façon, les observations sur le premier caractère viendront enrichir notre connaissance (et donc augmenteront la précision de nos estimations) des effets génétiques et de milieu pour le deuxième caractère et vice-versa. Ceci se fera malgré tout au prix de calculs nettement plus complexe. Là encore, avec le développement des matériels informatiques, la tendance actuelle est à l'utilisation judicieuse de modèles multicaractères.

Dans d'autres cas, notre modèle incluera d'autres effets aléatoires. Nous avons déjà donné l'exemple du modèle à effets maternels. Un autre exemple important est l'ensemble des modèles dits "à répétabilité" pour analyser des situations où on peut avoir des mesures répétées du même caractère sur un même animal (plusieurs lactations d'une même vache par exemple). On introduit alors un effet aléatoire "d'environnement permanent" qui traduit le fait que les performances d'un même animal se ressemblent plus (sont plus corrélées) que les performances de deux vrais jumeaux.

6 / Estimation des paramètres génétiques

L'étude génétique d'un caractère quantitatif et l'évaluation des reproducteurs nécessitent la connaissance de "paramètres génétiques" - héritabilité, corrélations génétiques, variances génétiques - qui sont tous fonctions de composantes de la variance totale σ_p^2 du caractère. Ainsi, lorsque nous avons décrit la valeur génétique additive et la résiduelle comme effets aléatoires d'un modèle animal, nous avons supposé que leur variance ($\sigma_a^2=h^2\sigma_p^2$ et $\sigma_e^2=(1-h^2)\sigma_p^2$) était connue. Pour les obtenir, on utilise généralement des méthodes d'"estimation des composantes de la variance" qu'on applique aux mêmes types de modèle linéaire que nous avons décrit jusqu'ici. Dans les cas les plus simples, ces méthodes sont dérivées des techniques d'analyse de variance : la part respective de la variabilité totale qui est due aux différences entre animaux (ou entre leurs parents) et à l'environnement est prise égale à son espérance, qui est une fonction simple de σ_a^2 et σ_e^2 (par exemple, la contribution d'un père à la variabilité totale mesurée sur ses descendants issus de mères non apparentées est égale à $(1/4)\sigma_a^2$). Mais dans la plupart des cas, on doit estimer ces composantes de la variance à partir de données déséquilibrées (avec des nombres très différents de descendants par père ou d'animaux par élevage par exemple, et avec des degrés de parenté très divers) et sélectionnées. On doit avoir recours à des méthodes beaucoup plus lourdes, à la fois d'un point de vue théorique et calculatoire. Celle donnant les résultats les plus satisfaisants - mais aussi les calculs les plus complexes - est le REML (Restricted Maximum Likelihood ou Maximum de Vraisemblance Restreinte) qui prend en compte la perte d'information liée à l'estimation simultanée des effets fixes mais aussi - dans le cas du modèle animal - les conséquences de la sélection et des accouplements non aléatoires sur l'évolution de la variance génétique.

Conclusion

La modélisation est une étape délicate dans l'évaluation des reproducteurs et l'estimation des paramètres génétiques, comme dans toute analyse de données. Le modèle comporte toujours 3 parties: l'équation du modèle, une description des caractéristiques statistiques des éléments du modèle (fixes ?, aléatoires ?, distribution ?, etc...) et les hypothèses supplémentaires qui sont faites (pas d'autres effets de milieu importants, ancêtres non apparentés et non sélectionnés, etc...). Tous ces aspects nécessitent une bonne connaissance des phénomènes biologiques et des conditions d'élevage et une bonne rigueur dans la formulation exhaustive des hypothèses. La validité des résultats ultérieurs et des conclusions qui seront tirées en dépend.

Références bibliographiques

Berger J.O., 1985. Statistical decision theory and Bayesian analysis. Springer-Verlag, New-York, 627 pp.

Bonaïti B., Boichard D., Verrier E., Ducrocq V., Barbat A., Briend M., 1990. La méthode française d'évaluation génétique des reproducteurs laitiers. Prod. Anim. INRA, 3, 83-92

Ducrocq V., 1990. Les techniques d'évaluation génétique des bovins laitiers. Prod. Anim., INRA, 3, 3-16

Hazel L.N., 1943. The genetic basis for constructing selection indexes. Genetics, 28, 476-490.

Henderson C.R., 1963. Selection index and expected genetic advance in Hanson W.D. et H.F. Robinson (Eds.). Statistical genetics and plant breeding pp 141-163. National Academy of Science. National Research Council, Publ. 982, Washington, DC.

Henderson C.R., 1973. Sire evaluation and genetic trend. in Proceeding of the Animal Breeding and Genetics Symposium in honor of Dr. J.L. Lush, Blacksburgh, Virgina. August 1972 pp 10-41.

Henderson C.R., 1984. Applications of linear models in animal breeding, 462 pp. University of Guelph. Canada.

Henderson C.R., 1988. Theoretical basis and computational methods for a number of different animal models. J. Dairy Sci., 71 (supp 2) 1-16.

Minvielle F., 1990. Principes d'amélioration génétique des animaux domestiques. INRA, Paris 211p.

Quaas R.L., 1984. Use of mixed model for Prediction in BLUP School handbook, R.L. Quaas, R.D. Anderson, A.R. Gilmour, pp 1-77. Animal Genetics and Breeding Unit. University of New England, Australie.

Sellier P., 1982. Selecting populations for use in crossbreeding. In Proceeding of the 2nd World Congress on Genetics applied to Livestock Production. Madrid, Espagne, Vol. VI, 15-49.

Smith H.F., 1936. A discriminant function for plant selection. Annals of Engenics, 7, 240-250.